visiting only site A. Thus, if a future campaign proposes to use sites A, B, and C (but not D), the projected reach from site A may be discounted by 5%, because 5% of users will have been reached on those other sites. In practical terms, the discount may effectively be one half of the projected overlap, since the same calculation for each of sites B and C, will properly compensate for the other half of the overlap users.

Step 204 entails the collection of data providing population size and demographic information on the various Internet sites under consideration for the advertising campaign. This is normally conducted by an outside Rating firm not shown in Figure 1, analogous to the firms that estimate television viewership. The population information collected indicates the total number of "hits" or potential impressions the site can generate in a given time period. Essentially, this measures the size of the advertiser's audience. Demographic information is also collected about the advertiser's audience. Because web users are anonymous, demographic information is collected through surveys and other conventional research tools, as with broadcast media ratings services. Demographic information may include age, sex, income, parental status, and geographic location, for instance.

The above information may be collected in any order, without one step being dependent on the next as illustrated. Once the information is collected, an advertising campaign is statistically simulated. Using the frequency distributions, the Monte Carlo method is preferably used to simulate a buy of a certain impression level on each of several selected sites.

The simulation proceeds for each site by segregating the users (i.e., cookies) recorded to have visited that site into groups, to generate "buckets" or "bins" of users. The users are sorted based on what the frequency data indicates is the expected number of impressions they have received in the past, with the most active users in the top decile, and the least active users in the bottom decile. A simplified example of this follows, in which the total user population is 100 cookies, and 1000 advertisements are to be served in the simulation:

Cookie numbers	Bin population	% of impressions served	Cumulative % of impressions	# of ads allocated	Ad ID numbers
1-10	10	50	50	500	1-500
11-20	10	25	75	250	501-750
21-30	10	15	90	150	751-900
31-40	10	5	95	50	901-950
41-50	10	2	97	20	951-970
51-60	10	1	98	10	971-980
61-70	10	..8	98.8	8	980-988
71-80	10	.06	99.4	6	988-994
81-90	10	.04	99.8	4	994-998
91-100	10	.02	100	2	998-1000

To run the simulation, each of 1000 advertisements are "served". First, the advertisement is assigned a random number in the range of the total number of ads (1-1000). Second, based on that number, it is assigned to the bin in which that ID number is found (e.g. if the ad is assigned number 635, it is assigned to the second bin associated with cookies 11-20.) Thus, the ad will be assigned to one of the cookies within that bin. Third, the ad is assigned to one of the cookie-members of the assigned bin by random choice. This proceeds with each of the advertisements simulated. After this, each cookie has been reached with a given number of advertisements, which is recorded and stored.

Those in each bin will likely have different numbers of ads assigned, as the randomizing effects creates a statistical distribution within each bin. Some members of lower bins may receive more ads than some members of relatively higher bins. However, because this randomizing effect is based on actual probabilities, and not simple statistical noise, a smoother and more useful distribution will be achieved in the result, which will show that a certain number received zero ads, another number received one ad, another number received two, etc. For each integer number of ads that may have been served, a certain number of the cookies received that number. This data may usefully be converted into a simpler form, by stating that x percent of cookies received an advertisement, or y percent were reached by at least n advertisements. Alternatively, a useful form of to display the results in what is known as a

frequency histogram. This is a summary table indicating how many cookies received *n* impressions, for every integer *n* up to a certain point.

This is preferable to a non-randomized scheme, in which the advertisements are presumed smoothly distribute (with exactly 5 ads being served to each of cookies 1-10, for instance, and exactly 2.5 to each of cookies 11-20.) This creates a stepped, discontinuous result, that introduces thresholds that do not exist in reality, in addition to the problem of fractional ads. The chief limitation of the completely deterministic process is that the frequency histogram is not smooth. This lack of smoothness becomes a problem when one wishes to view Effective Reach for consecutive frequencies. A small change in frequency (say from 2 to 3) can produce a sharp change in the number of cookies. This contradicts the behavior observed empirically in actual campaigns where the frequency histogram describes a smooth curve (over the discrete set of integers). The disclosed method of prediction much more closely predicts eventual results than do prior methods.

A potential drawback for the Monte Carlo bucketing method is that it can be computationally costly to run. In particular, an application that sat on a users desktop could take a prohibitively long time to accomplish the estimation. Therefore, an additional technique must be used to process the output of the Monte Carlo method. For every Site and for Effective Frequencies from 1 to 15, the Effective Reach for many impression levels is calculated. For each frequency level, a series of points are produced. these points describe the interplay between reach and impressions under the Monte Carlo method. Moreover, these points describe a smooth curve. One may fit curves that describe the relationship between Impressions and Effective Reach for each of the effective frequencies from 1 to 15. This process can be run intermittently, then those curves can be evaluated by the application in real time to produce frequency estimates.

Having converted the recorded frequency data into a useful form, the impression levels for a proposed Media Plan may then be input in step 214, and converted into the desired information.